Optimization Algorithms for Parameter Estimation and Data Reconciliation

L. T. Biegler
Chemical Engineering Department
Carnegie Mellon University
Pittsburgh, PA

Parameter Estimation and Data Reconciliation

I. Motivation and Formulation
- Problem Formulation
- Derivation of Objectives

II. Unconstrained Problems
- Linear Least Squares
- Gauss-Newton Methods
- Trust Region Methods

III. Constrained Least Squares
- Problem Formulation
- Specialization to SQP
- EVM Problems

IV. Statistical Inference
- Linear Problems
- Nonlinear Problems

V. Introduction to Data Reconciliation
- Optimization Problem
- Gross Error Detection

VI. Contaminated Normal
- Derivation
- Examples

VII. Robust Statistics (M-Estimator)
- Huber (Fair function)
- Hampl (Redescending function)
- Examples

VIII. Mixed Integer Approaches
- Formulations
- Comparisons

2

Parameter Estimation References

Bard, Y., Nonlinear Parameter Estimation, Academic Press, New York (1974)

Boggs, P. T., R. H. Byrd, And R. B. Schnabel (1987), "A Stable and Efficient Algorithm For Nonlinear Orthogonal Distance Regression," SIAM J. Sci. Stat. Comput., 8(6):1052-1078. (ODRPACK)

Caracotsios, M., PhD Thesis, University of Wisconsin (1985) (GREG)

Dennis, J.E., Gay, D.M., and Welsch, R.E. (1981), *ACM Trans. Math. Software*, Vol. 7, No. 3. (NL2SOL)

Lawson, C.L., and Hanson, R.J. (1974), **Solving Least Squares Problems**, Prentice-Hall, Englewood Cliffs, N.J.

More, J.J. (1978), *Lecture Notes In Mathematics No. 630*, G.A. Watson (ed.), Springer-Verlag, Berlin (MINPACK)

Nocedal, J. and S. Wright, Numerical Optimization, Springer Verlag, Berlin (1998)

Tjoa, I-B and L.T. Biegler, "Simultaneous Solution and Optimization Strategies for Parameter Estimation of Differential-Algebraic Equation Systems," I&EC Research, 30, p. 376 (1991).

Tjoa, I-B and L. T. Biegler, "A Reduced Successive Quadratic Programming Strategy for Errors-in-Variables Estimation," Computers and Chemical Engineering 16, 6, p. 523 (1992)

3

Data Reconciliation References

Texts:

Shankar Narasimhan, Cornelius Jordache, Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data, Gulf Pub Co, 2000

J.A. Romagnoli, M. Sanchez, Data processing and reconciliation for chemical process operation, Academic Press International, ISBN 0-12-594460-8 (2000).

Papers:

Tjoa, I-B and L.T. Biegler, "Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems," Computers and Chemical Engineering 15, 10, p. 679 (1991)

Albuquerque, J. S., and L. T. Biegler, "Gross Error Detection and Variable Classification in Dynamic Systems," AIChE Journal, 42, 10, p. 2841 (1996)

Arora, N. and L. T. Biegler, "Redescending estimators for data reconciliation and parameter estimation," Computers and Chemical Engineering, 25, p. 1585 (2001)

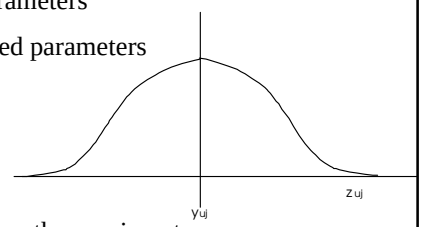
and citations therein

4

Maximum Likelihood Derivation - 1

Motivation:

- fit models to data to find 'optimal' parameters
- determine levels of confidence of fitted parameters
- evaluate suitability of models



Derivation of Objective Function

Let $z_{uj} = y_{uj}(\theta) + \varepsilon_{uj}$

z_{uj} - jth component of data for uth experiment

y_{uj} - corresponding (correct) model value

ε_{uj} - observation error following probability distribution function (pdf), p(z)

θ - adjustable parameters in model

What is p(z)? For a scalar z, how is it derived?

5

Maximum Likelihood Derivation - 2

Assume the following moment information:

$$p(z) \geq 0, \int_{-\infty}^{\infty} p(z)dz = 1, \int_{-\infty}^{\infty} z\ p(z)dz = \eta, \int_{-\infty}^{\infty} (z-\eta)^2\ p(z)dz = \sigma^2$$

Define measure of information (Shannon, 1948) and find the distribution function that maximizes the information assuming only the moment information:

$$Max\ I(p) = E(\log p) = \int_{-\infty}^{\infty} \log(p(z))\, p(z)dz$$

$$s.t.\ p(z) \geq 0,\ \int_{-\infty}^{\infty} p(z)dz = 1$$

$$\int_{-\infty}^{\infty} z\ p(z)dz = \eta,\ \int_{-\infty}^{\infty} (z-\eta)^2\ p(z)dz = \sigma^2$$

Problem can be solved analytically to yield: $p(z) = \frac{1}{(\sqrt{2\pi})\sigma} \exp(-(z-\eta)^2/(2\sigma^2))$

If z is an m-vector for a single experiment, this can be extended to a joint multivariable distribution to give:

$$p(z) = (2\pi)^{-m/2} \det(V)^{-1/2} \exp(-1/2(z-\eta)^T (V)^{-1}(z-\eta))$$

where V is a covariance matrix defined by: $V = \int (z-\eta)(z-\eta)^T\ p(z)dz$

6

Maximum Likelihood Derivation - 3

Consider n multiple experiments (with index u). Each experiment has a mean η_u, covariance V_u and experimental error distribution, ε_u

$$p(\varepsilon_u) = (2\pi)^{-m/2} \det(V)^{-1/2} \exp(-1/2(\varepsilon_u)^T (V)^{-1}(\varepsilon_u))$$

and the joint probability distribution is given by:

$$\prod_{u=1}^{n} p(\varepsilon_u) = (2\pi)^{-mn/2} \left[\prod_{u=1}^{n} \det(V_u)^{-1/2} \right] \exp(-1/2 \sum_{u=1}^{n} (\varepsilon_u)^T (V_u)^{-1}(\varepsilon_u))$$

This distribution now needs to be converted into an objective function that 'maximizes information' about our data.

Let's make the following assumptions:

- Replace distributional errors, ε_u by the actual residuals, $e_u = (z_u - y_u(\theta))$
- Experiments u are independent and V_u is the same for all experiments, $V = E(\varepsilon_u \varepsilon_u^T)$
- Define likelihood function $L(\theta) = \Pi\, p(e_u)$, and maximize this function (or its log).

This leads to the general form:

$$\log L(\theta) = -(nm/2)\log(2\pi) - (n/2)\log(\det(V)) - 1/2 \sum_{u=1}^{n} (e_u)^T (V)^{-1}(e_u)$$

7

Maximum Likelihood Objective Functions

Specialize objectives based on what we know about the error distributions.

Define moment matrix:

$$M(\theta) = \sum_{u=1}^{n} (e_u)(e_u)^T, \; Tr(V^{-1}M(\theta)) = \sum_{u=1}^{n} (e_u)^T (V)^{-1}(e_u)$$

and $$\log L(\theta) = -(nm/2)\log(2\pi) - (n/2)\log(\det(V)) - 1/2 \sum_{u=1}^{n} Tr(V^{-1}M(\theta))$$

Since first two terms do not contain θ, we simply minimize $Tr(V^{-1} M(\theta))$

Special cases:

- <u>Ordinary Least Squares:</u> V is known, all component errors e_{uj} have same distribution and are independent of each other, i.e., V = v I

$$\text{Min } Tr(M(\theta)) = \Sigma_u \Sigma_j e_{uj}^2$$

- <u>Simple Weighted Least Squares:</u> V is known and diagonal, all component errors e_{uj} are independent of each other, i.e., $V = \text{diag}\{\sigma_j^2\}$

$$\text{Min } Tr(V^{-1}M(\theta)) = \Sigma_u \Sigma_j e_{uj}^2/\sigma_j^2$$

- <u>Weighted Least Squares:</u> V is known but general, all component errors e_{uj} depend on each other:

$$\text{Min } Tr(V^{-1}M(\theta)) = \Sigma_u \Sigma_j e_u^T V^{-1} e_u$$

8

Maximum Likelihood – Unknown Covariance

Maximize L(θ) wrt V as well as θ.

$$\log L(\theta) = -(nm/2)\log(2\pi) - (n/2)\log(\det(V)) - 1/2\sum_{u=1}^{n} Tr(V^{-1}M(\theta))$$

Assuming V is symmetric and nonsingular, we get:

$$\frac{\partial \log L(\theta)}{\partial V} = -(n/2)V^{-1} - 1/2\sum_{u=1}^{n}(V^{-1}M(\theta)V^{-1}) = 0$$

which leads to: $V^* = 1/n\, M(\theta)$

Substitution into L(θ) leads to:

$$\log L(\theta) = -(nm/2)\log(2\pi) - (n/2)\log(n) - 1/2mn - n/2\log\det(M(\theta))$$

so we minimize: $n/2 \log \det(M(\theta))$

Notes:

- If the structure of V is known, we can further specialize the objective for unknown covariance (e.g., diagonal covariance, same covariance)
- V* is biased but can be corrected by using V" = n/(n-m n_θ) V*

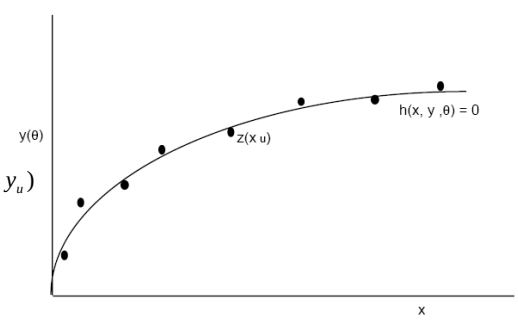
Unconstrained Least Squares

Basic Problem:

$$Min\ \Phi(\theta, y, z) = 1/2\sum_{u=1}^{n}(z_u - y_u)^T W_u (z_u - y_u)$$

$$s.t.\ \ h_u(x_u, y_u, \theta) = 0$$

$$a \le \theta \le b$$



- W_u can be chosen to be $(V_u)^{-1}$
- $h_u(x_u, y_u, \theta) = 0$ is a model of the system
- x_u are fixed independent variables
- Bounds are not expected to be active

Consider the case where y_u is an explicit function of θ,

$$h_u(x_u, y_u, \theta) = 0 \implies y_u = f_u(x_u, \theta)$$

and $$Min\ \Phi(\theta, y, z) = 1/2\sum_{u=1}^{n}(z_u - f_u(\theta))^T W_u (z_u - f_u(\theta))$$

Least Squares – Solution Methods

Linear Least Squares

Model is given by: $y_u = A_u\theta + b_u$

$$Min\ \Phi(\theta, z) = 1/2\sum_{u=1}^{n}(z_u\text{-}A_u\theta - b_u)^T W_u (z_u\text{-}A_u\theta - b_u)$$

From $\nabla_\theta \Phi(\theta, z) = 0$, we get the normal equation:

$$\sum_{u=1}^{n}(A_u{}^T W_u A_u)\theta = \sum_{u=1}^{n} A_u{}^T W_u (z_u - b_u)$$

and this leads to: $\theta = \left(\sum_{u=1}^{n}(A_u{}^T W_u A_u)\right)^{-1}\sum_{u=1}^{n} A_u{}^T W_u (z_u - b_u)$

Note: A better way to solve this linear system is to do a QR factorization on a concatenation (over u) of $(Wu)^{1/2}A_u$.

Least Squares – Solution Methods

Nonlinear Least Squares

Model is given by: $y_u = f_u(x_u\theta)$

$$Min\ \Phi(\theta, z) = 1/2\sum_{u=1}^{n}(z_u\text{-}f_u(\theta))^T W_u (z_u\text{-}f_u(\theta))$$

And we solve this unconstrained problem with Newton's method

From Taylor series expansion: $\nabla_\theta \Phi(\theta^k, z) + \nabla_{\theta\theta}\Phi(\theta^k, z)\Delta\theta = 0$

$$\nabla_\theta \Phi(\theta^k, z) = -\sum_{u=1}^{n} J_u^T W_u (z_u - f_u(\theta^k)),\ J_u = \nabla_\theta f_u{}^T$$

$$\nabla_{\theta\theta}\Phi(\theta^k, z) = \sum_{u=1}^{n}(J_u^T W_u J_u + R_u),\ \{R_u\}_{ij} = -\nabla_{\theta_i\theta_j} f_u^T W_u (z_u - f_u(\theta^k))$$

Now assume that $(z_u\text{-}f_u(\theta))$ is nearly zero and therefore R_u is nearly zero. Then the Hessian simplifies to: $\nabla_{\theta\theta}\Phi(\theta^k, z) = \sum_{u=1}^{n} J_u^T W_u J_u$

and we have the **Gauss-Newton Method**:

$$\Delta\theta = \left(\sum_{u=1}^{n}(J_u{}^T W_u J_u)\right)^{-1}\sum_{u=1}^{n} J_u{}^T W_u (z_u - f_u(\theta^k))$$

Note: A better way to solve this linear system is to do a QR factorization on a concatenation (over u) of $(W_u)^{1/2}J_u$.

Globalization of Gauss-Newton Method

To ensure convergence from poor starting points:

<u>Line search method</u>

Choose $\alpha \in (0, 1]$ so that a sufficient decrease is found for $\Phi(\theta)$ with: $\theta^{k+1} = \theta^k + \alpha\, \Delta\theta$.

This will converge to a stationary point ($\nabla_\theta \Phi = 0$) as long as

$\sum_{u=1}^{n} J_u^T W_u J_u$ is sufficiently positive definite. What if singular?

Add λ I to Hessian to get the **<u>Levenberg-Marquardt method</u>**.

$$\Delta\theta = \left(\sum_{u=1}^{n} (J_u{}^T W_u J_u) + \lambda I \right)^{-1} \sum_{u=1}^{n} J_u{}^T W_u (z_u - f_u(\theta^k))$$

How should λ be adjusted?

13

Unconstrained Problems: TR Motivation

Take: $Min \quad \Phi(\theta)$

Model Problem: $Min \quad \Phi(\theta_k) + \nabla\Phi(\theta_k)^T \Delta\theta + \frac{1}{2}\Delta\theta \nabla^2\Phi(\theta_k)^T \Delta\theta$

$$\frac{1}{2}\|\Delta\theta\|^2 \le \frac{1}{2}\Delta^2$$

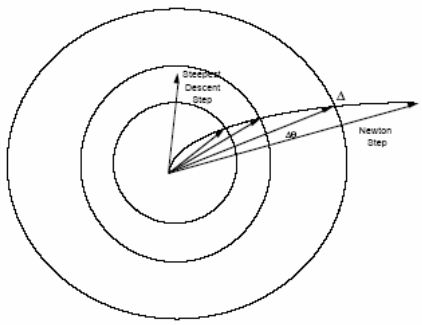
Optimality Conditions: Leads to Levenberg-Marquardt Method

$$\left\{ \begin{array}{c} \nabla\Phi(\theta_k) + \nabla^2\Phi(\theta_k)^T \Delta\theta + \lambda\Delta\theta = 0 \\ \|\Delta\theta\|^2 \le \Delta^2, \lambda \ge 0 \\ \lambda(\|\Delta\theta\|^2 - \Delta^2)/2 = 0 \end{array} \right\} \Longrightarrow \Delta\theta = -(\nabla^2\Phi(\theta_k) + \lambda I)^{-1} \nabla\Phi(\theta_k)$$

How do we choose λ or Δ?

14

Extremes of Trust Region Method



For given Δ, solve for λ directly

$$1/||\Delta\theta(\lambda)|| - 1/\Delta = 0$$

$\lambda = 0$, Δ large: $\Delta\theta$ is the Gauss-Newton Step

$\lambda \dashrightarrow \infty$, $\Delta = 0$: $\Delta\theta$ is a small step in the steepest descent direction

Trust Region Methods - guaranteed to converge to local optima, with much weaker assumptions than line search methods.

15

Trust region approach in MINPACK (More', 1980)

Choose λ so that $\| \Delta\theta \| \leq \Delta$, comparing τ = ared/pred

- actual reduction (ared): $\Phi(\theta^k) - \Phi(\theta^k + \Delta\theta)$
- predicted reduction (pred): $\nabla_{\theta\theta}\Phi^T\Delta\theta + 1/2\, \Delta\theta^T\nabla_{\theta\theta}\Phi\, \Delta\theta$

At iteration k:

- Calculate λ corresponding to Δ, and calculate $\Delta\theta$.
- Evaluate (ared) and (pred).
- Define τ = ared/pred

  a) If $\rho_1 < \tau \leq \rho_0$, $\Delta = m_1 \Delta$

  b) If $\rho_2 \leq \tau < \rho_1$ or $\rho_0 < \tau$, $\Delta = \Delta$

  b) If $\tau < \rho_2$, $\Delta = \Delta / m_1$

  c) If $\tau < \rho_3$ reset $\Delta\theta = 0$

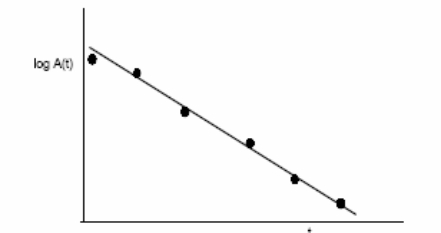
  Set, $\theta^{k+1} = \theta^k + \Delta\theta$

Typical values for the parameters:

$m_1 = m_2 = 2$ and $\rho_1 = 0.75$, $\rho_2 = 0.25$, $\rho_3 = 0$, $\rho_0 = 2$.

16

Why is trust region required for parameter estimation?



First order reaction: $A(t) = A(0) \exp(-k\, t)$; $k = k_0 \exp(-E/RT)$

Data available only at one temperature.

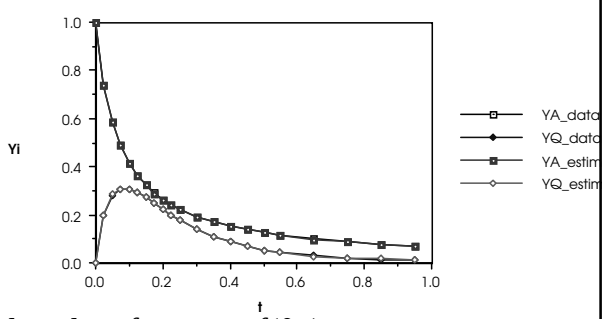
<u>Results:</u>

- Nonunique parameter estimates, k_0, E
- Singular Hessian
- Can be due to poor model and/or poor data
- Trust region methods will ensure convergence to "some" solution,
- Postoptimality analysis will establish nonuniqueness, insensitivity

Example: Catalytic Cracking of Gasoil (Tjoa, 1991)

$$A \xrightarrow{k_1} Q \xrightarrow{k_2} S, \quad A \xrightarrow{k_3} S$$

$$y_A' = -(k_1+k_3)\, y_A^2$$
$$y_Q' = -k_1 y_A^2 - k_2\, y_Q$$
$$y_A(0) = 1,\ y_Q(0) = 0$$



ODEs can be solved to yield explicit form: $y = f(\theta,t)$

Apply Trust Region method (GREG):

$(k_1, k_2, k_3)_0 = (6, 4, 1)$

$(k_1, k_2, k_3)_* = (11.95, 7.99, 2.02)$

$(k_1, k_2, k_3)_{true} = (12, 8, 2)$

Converges in 5 iterations (11 function calls)

Convergence Rates for Gauss-Newton Methods

Small Residuals at Solution (Good Model Fit):

$e_u^* = 0$, $R_u^* = 0$,

$$\nabla_{\theta\theta}\Phi(\theta^*,z) = \sum_{u=1}^{n} J_u^T W_u J_u$$

- Gauss-Newton method is quadratically convergent.
- Trust region will be inactive if Hessian is nonsingular
- L-M is also quadratically convergent for unique θ^*.

Large Residuals at Solution (Poor Model Fit):

$e_u^* \neq 0$, $R_u^* \neq 0$,

$$\nabla_{\theta\theta}\Phi(\theta^*,z) = \sum_{u=1}^{n} J_u^T W_u J_u + R_u$$

- Gauss-Newton method is linearly convergent.
- Trust region may not be inactive if Hessian is nonsingular
- L-M is also linearly convergent for unique θ^*.

19

Hybrid Methods (Gauss and quasi-Newton)

Quasi-Newton Methods

- DFP and BFGS Methods apply secant formula, symmetry and positive definiteness of Hessian
- Do not exploit structure of least squares problem

Dennis, Gay and Welsh (1981) - NL2SOL

- Approximates true Hessian as G-N Hessian is known
- Specialized, self-scaling Q-N method developed that approximates R_u
- Incorporates Trust Region Approach of More'
- Leads to superlinear convergence

Fletcher and Xu (1987)

- Applies specialized Q-N method to approximate R_u
- Uses a switching rule to monitor if there are small or large residuals

$$\tau^k = (\Phi(\theta^k) - \Phi(\theta^{k+1}))/\Phi(\theta^k)$$

Large residuals: $\lim_{k\to\infty} \tau^k = 0$, Use specialized Q-N update if $\tau^k \leq \varepsilon$

Small residuals: $\lim_{k\to\infty} \tau^k = 1$, Use specialized G-N formula if $\tau^k > \varepsilon$

(Choose $\varepsilon \sim 0.1$)

20

Constrained Least Squares

Motivation:

- Model cannot be reformulated as $y = f(\theta)$
- Too expensive to converge $h_u(y_u, \theta) = 0$ for each parameter value

Basic Formulation:

$$Min\ \Phi(\theta, y, z) = 1/2 \sum_{u=1}^{n} (z_u - y_u)^T W_u (z_u - y_u)$$
$$s.t.\ \ h_u(y_u, \theta) = 0$$
$$a \le \theta \le b$$

- Any NLP method can be used to solve this problem
- SQP can be tailored to take advantage of special form of Φ allows for tailored algorithm.
- Leads to faster algorithm than standard SQP with BFGS updates

21

Optimization Strategy: SQP method

Let $x^T = [\theta^T, y^T]$ and consider QP subproblems for SQP:

$$\min \quad \nabla\Phi(x^k)^T d + 1/2\, d^T B\, d$$
$$\text{s.t. } h(x^k) + \nabla h(x^k)^T d = 0$$
$$x^L \le x^k + d \le x^U$$

First order necessary conditions

$$\begin{bmatrix} B & \nabla h \\ \nabla h^T & 0 \end{bmatrix} \begin{bmatrix} d \\ v \end{bmatrix} = - \begin{bmatrix} \nabla\Phi \\ h \end{bmatrix}$$

Problems:

- How to deal with a larger QP problems
- How to calculate the Hessian

Strategies:

- Use Range and Null space Decomposition strategy to decompose the search direction into:
  - Null space movement
  - Range space movement
- Use a hybrid Gauss-Newton and BGFS update formula
- Analogy to unconstrained approaches

22

Range and Null Space Decomposition

Define linear QP system as: M s = f, to give:

$$\begin{bmatrix} B & \nabla h \\ \nabla h^T & 0 \end{bmatrix} \begin{bmatrix} d \\ v \end{bmatrix} = - \begin{bmatrix} \nabla \Phi \\ h \end{bmatrix}$$

and select an n x n nonsingular matrix: H = [Y | Z], where $\nabla h^T Z = 0$.

- Z, Y are null & range space bases for the linearized equalities
- Search direction with range (p_Y) and null space (p_Z) components:

$$d = Y p_Y + Z p_Z. \quad Y^T = [\ 0 \mid I\ ] \qquad Z^T = [\ I \mid N^T C^{-T}]$$

Now defining X = diag [[Y | Z] , I], we can consider the equivalent system $X^T M X z = X^T f$ (with X z = s) as:

$$\begin{bmatrix} Y^T BY & Y^T BY & Y^T \nabla h \\ Z^T BY & Z^T BZ & 0 \\ \nabla h^T Y & 0 & 0 \end{bmatrix} \begin{bmatrix} p_Y \\ p_Z \\ v \end{bmatrix} = - \begin{bmatrix} Y^T \nabla \Phi \\ Z^T \nabla \Phi \\ h \end{bmatrix}$$

Standard assumptions: set $Y^T B\, Y = 0$ and $Y^T B\, Z = 0$

23

Structure of Least Squares Hessian

$$B^k = \begin{bmatrix} \nabla_{\theta\theta} L & \nabla_{\theta y} L \\ \nabla_{y\theta} L & \nabla_{yy} L \end{bmatrix}$$

$$\nabla_{\theta\theta} L = \sum_{j=1}^{m} v_j \nabla_{\theta\theta} h_j \qquad \nabla_{\theta y} L = \sum_{j=1}^{m} v_j \nabla_{\theta y} h_j$$

$$\nabla_{y\theta} L = \sum_{j=1}^{m} v_j \nabla_{y\theta} h_j \qquad \nabla_{yy} L = \nabla_{yy} \Phi + \sum_{j=1}^{m} v_j \nabla_{yy} h_j$$

where $L(\theta, y, v) = \Phi(y) + v^T h(\theta, y)$

KKT multipliers (based on first order estimates) are given by:

$$v = - (Y^T \nabla h)^{-1} Y^T \nabla \Phi = - (Y^T \nabla h)^{-1} Y^T\ [\ 0 \mid \Sigma_u e_u^T W_u]^T$$

Assumption: If the residuals are *small*, then at convergence $e_u \approx 0 \Rightarrow v \approx 0$

The Hessian becomes $B^{G\text{-}N} = \begin{bmatrix} 0 & 0 \\ 0 & (W_u) \end{bmatrix}$

$\Rightarrow$ *Newton-like* rate of convergence

24

Least Squares Hybrid SQP Method

Motivation: Choose best Hessian approximation for different problem types

Strategies: Develop a switching rule to decide if Q-N or G-N approximation should be made for B^k. (Fletcher and Xu, 1987)

- Define merit function: $L^*(x^k) = \Phi(x^k) + v^T h(x^k) + 1/2\, \gamma\, \|h(x^k)\|^2$
- Uses a switching rule to monitor if there are small or large residuals

$$\tau^k = (L^*(x^k) - L^*(x^{k+1}))/L^*(x^k)$$

Large residuals: $\lim_{k \to \infty} \tau_k = 0$, Use specialized Q-N for $Z^T B Z$ if $\tau_k \leq \varepsilon$

Small residuals: $\lim_{k \to \infty} \tau_k = 1$, Use $Z^T B^{GN} Z$, $Z^T B^{GN} Y p_Y$ formula, $\tau_k > \varepsilon$

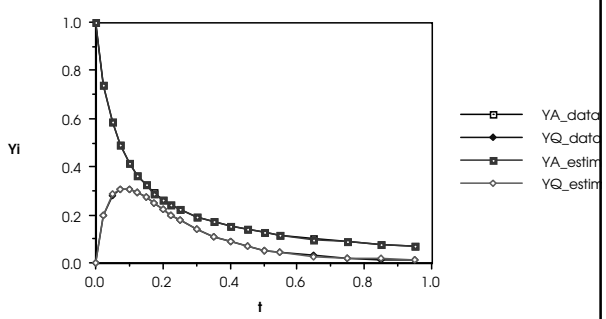
Choose $\varepsilon \sim 0.2$

25

Example: Catalytic Cracking of Gasoil (Tjoa, 1991)

$$A \xrightarrow{k_1} Q \xrightarrow{k_2} S, \quad A \xrightarrow{k_3} S$$

$$y_A' = -(k_1+k_3)\, y_A^2$$
$$y_Q' = -k_1 y_A^2 - k_2\, y_Q$$
$$y_A(0) = 1,\ y_Q(0) = 0$$



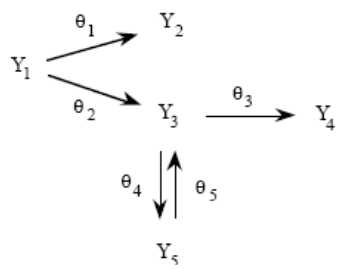
number of ODEs: 2
number of parameters: 3
discretized ODEs: 68 variables

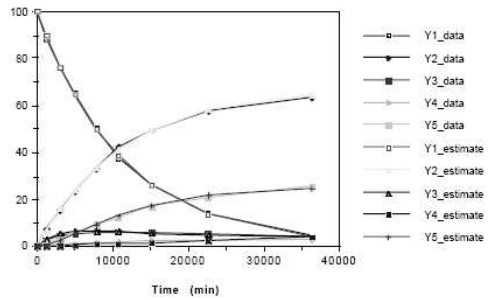
$(k_1, k_2, k_3)_0 = (6, 4, 1)$
$(k_1, k_2, k_3)_* = (11.95, 7.99, 2.02)$
$(k_1, k_2, k_3)_{true} = (12, 8, 2)$

Method	Obj.	Iters.	CPU (s, V3200)
BFGS SQP	8.23e-5	10	4.31
Gauss-Newton	8.23e-5	4	2.25
Hybrid SQP	8.23e-5	4	2.31

26

Small Residual Example: α-Pinene Kinetics





number of ODEs: 5
number of parameters: 5
discretized ODEs: 245 variables

Method	Iters.	CPU (s, V3200)
BFGS SQP	37	90.8
MINOS	21	64.8
Gauss-Newton	6	23.5
Hybrid SQP	6	23.6

27

Further Results (Tjoa and Biegler,1990)

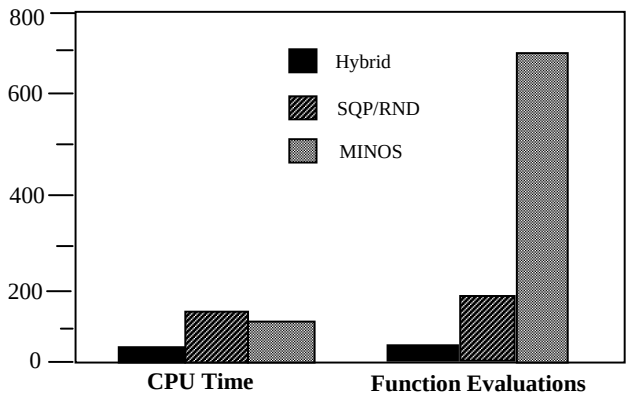
Number of Iterations (Function Evaluations)

Problem	GREG	MINOS	SQP based methods		
			BFGS	GN	Hybrid
1	6 (12)	8 (41)	8 (10)	3 (3)	3 (3)
2	14 (34	8 (72)	14 (17)	3 (3)	3 (3)
3	6 (12)	8 (55)	30 (31)	3 (3)	3 (3)
4	7 (20)	13 (95)	14 (15)	6 (6)	6 (6)
5	7 (19)	15 (236)	fail	13 (16)	13 (16)
6	5 (10)*	7 (72)	13 (21)	5 (5)	5 (5)
7	8 (16)	10 (150)	24 (30)	5 (5)	6 (6)
8	fail	21 (271)	37 (61)	6 (6)	6 (6)
9	26 (95)*	33 (320)	19 (24)	fail	18 (18)

*: using least squares objective function

28

Comparison of Hybrid vs. General Purpose NLP Codes (Tjoa)



- Summary on 10 parameter estimation (kinetics) problems
- Few parameters, degrees of freedom
- Hybrid method for Hessian structure:
  (Fletcher and Xu) quasi-Newton method.

29

Further Comparison – Constrained Trust Region Method (Arora, 2003)

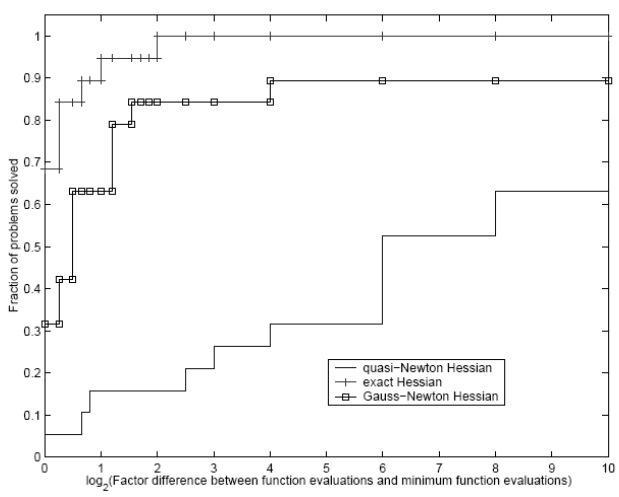


Figure 4.8: Parameter Estimation Problems

30

Statistical Inference of Estimated Parameters

Covariance of Optimal Parameters

Given an error distribution for the data (assumed Gaussian with covariance, V_z)
How does this affect the accuracy of the estimated parameters?

Recall: $V_z = E(\delta z\ \delta z^T)$ *and* $V\theta = E(\delta\theta^*\ \delta\theta^{*T})$

How does θ change with data z?

$$i.e.,\ \frac{\partial\Phi(\theta^*,z)}{\partial\theta}=0,\quad \frac{\partial\Phi(\theta^*+\delta\theta,z+\delta z)}{\partial\theta}=0$$

$$Approximate\ by: \frac{\partial\Phi^2(\theta^*,z)}{\partial\theta\,\partial z}\delta z+\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta^2}\delta\theta=0$$

$$and\quad \delta\theta=\left[\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta^2}\right]^{-1}\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta\,\partial z}\delta z,\ so\ we\ have:$$

$$V_\theta=E(\delta\theta\,\delta\theta^T)=\left[\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta^2}\right]^{-1}\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta\,\partial z}E(\delta z\,\delta z^T)\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta\,\partial z}^T\left[\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta^2}\right]^{-T}$$

$$\Rightarrow V_\theta=\left[\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta^2}\right]^{-1}\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta\,\partial z}V_z\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta\,\partial z}^T\left[\frac{\partial\Phi^2(\theta^*,z)}{\partial\theta^2}\right]^{-T}$$

31

Special Cases for Covariance

1. If the objective function has covariance independent of u, i.e.:

$$\Phi=\sum_{u=1}^{n}(z_u-f_u(\theta))^TV_z^{-1}(z_u-f_u(\theta))\ \text{ then we have: } V_\theta=\sum_{u=1}^{n}(J_u{}^TV_z^{-1}J_u)^{-1}$$

2. If z_u is a scalar, then V_z is σ^2 and V_θ is given by: $\sigma^2\sum_{u=1}^{n}\left(\frac{\partial f_u}{\partial\theta}\frac{\partial f_u}{\partial\theta}^T\right)^{-1}$

3. If the covariance for z is unknown, then estimate from moment matrix $V_z = M(\theta)/n$ and then: $V_\theta=\sum_{u=1}^{n}(J_u{}^TV_z^{-1}J_u)^{-1}$

4. For general likelihood functions, $V_\theta = -(\nabla_{\theta\theta}(\log L(\theta^*))^{-1}$ is asymptotically correct as $n \rightarrow \infty$.

32

Elliptical Confidence Regions

Single parameter

For a given interval let $\gamma = \Pr(a \le \theta^* \le b) = \int_a^b p(\theta^*|\theta_{true})\, d\theta$

and for a single parameter this becomes: $|\theta_{true} - \theta^*| \le \zeta$ where ζ is the confidence level for γ with σ_θ calculated with n --> ∞ and

$$\sigma_\theta^2 = \sigma^2 \sum_{u=1}^{n} \left(\frac{\partial f_u}{\partial \theta} \frac{\partial f_u}{\partial \theta}^T \right)^{-1}$$

Otherwise, with a small sample size: $|\theta_{true} - \theta^*| \le t\, s_\theta$

$$s = \frac{1}{n-1}\left[\sum_{u=1}^{n}(z_u - f_u(\theta))^2\right]^{-1} \qquad s_\theta^2 = s^2 \sum_{u=1}^{n} \left(\frac{\partial f_u}{\partial \theta} \frac{\partial f_u}{\partial \theta}^T \right)^{-1}$$

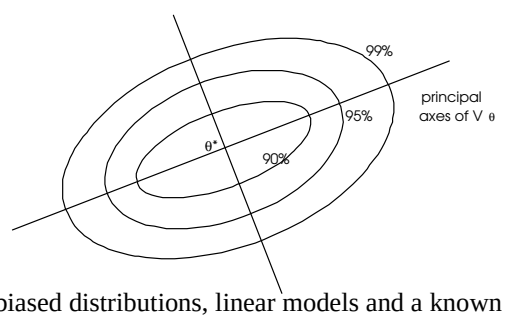
Multiple parameters

Map out a region $S(\theta)$ so that $\gamma = \Pr(\theta_{true} \in S(\theta))$

This can be done using the principal directions of V_θ which leads to:

$$\gamma = \Pr((\theta_{true}-\theta^*)^T V_\theta^{-1} (\theta_{true}-\theta^*))$$

33



For normal, unbiased distributions, linear models and a known V_θ, this probability follows a χ^2 distribution so that the region can be defined by:

$$(\theta_{true}-\theta^*)^T V_\theta^{-1} (\theta_{true}-\theta^*) \le c(\gamma)$$

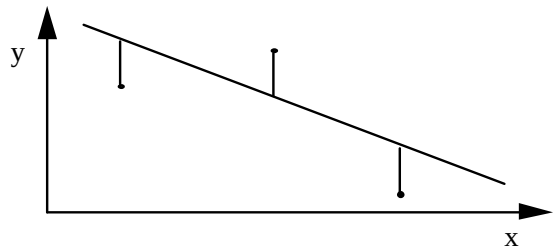
$c(\gamma)$ is χ^2 value for γ level of confidence with n_θ degrees of freedom.

- For a scalar z, the χ^2 test simplifies to an F-test for determination of $c(\gamma)$.
- Elliptical confidence regions are correct if the model is linear or for small levels of confidence, γ. Otherwise, confidence regions can deviate greatly from ellipses.
- Elliptical confidence regions are most commonly used. Nonlinear confidence regions much more expensive to calculate.

34

Errors in Variables Models (EVM)

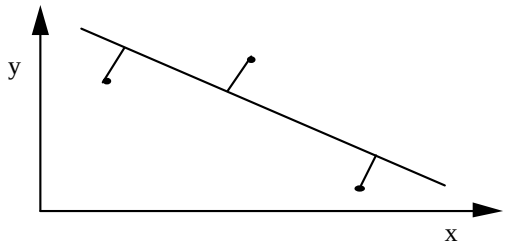
Conventional model: $y = f(x,\theta)$ or $h(x, y, \theta) = 0$



No. of independent variables (x) = s_1
No. of dependent variables (y) = s_2
No. of parameters (θ) = p
No. of constraints = m
No. of data sets = r
$\Rightarrow$ *No error in the "independent" variables, x*
Number of degrees of freedom for optimization = p
Minimize vertical distance, e.g. $(y_u - z_u)^2$

Errors in Variables Models (EVM) - 2

Implicit Model: $f(x,y,\theta) = 0$
Both x and y have inherent measurement errors
$\Rightarrow$under-determined system
(e.g., pressure vs. temperature data)



Number of degrees of freedom for optimization
= p + (s-m)r
$\Rightarrow$ NLP size grows linearly with the number of data sets
Minimize a *nonvertical* distance

EVM Problem Statement

$$Min\ \ \Phi = 1/2\sum_{u=1}^{r} e_u^T W e_u$$
$$s.t.\ \ f_u(w_u, \theta) = 0$$
$$a \le \theta \le b$$
where $e_u = w_u - z_u,\ w_u^T = [x_u^T\ y_u^T]$

$$Min\ \ \Phi = \sum_{u=1}^{r} \phi_u = 1/2\sum_{u=1}^{r} e_u^T W e_u$$
$$s.t.\ \ h_u(w_u, \zeta_u, \theta) = 0$$
$$a \le \theta \le b$$
where $h_u^T = [\,f_u^T\ g_u^T\,]$

Formulation:
- Least squares nonlinear constraints with many degrees of freedom

Current Approaches
- Nonlinear Programming Strategies – Expensive if sparsity not exploited
- Linearized Least Squares - Not robust, global convergence not enforced

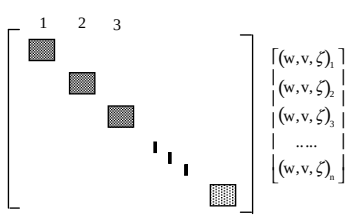
Introduce
New variables: ζ_u
New constraints: $g_u = \zeta_u - \theta = 0$.

Solution Strategy:
- Apply decoupling strategy to SQP method
- Decompose for each data set
- Computational cost linear in # of data sets
- Can solve in parallel

37

KKT Conditions for EVM

Structure of the KKT matrix after decoupling:



= RHS(Δθ)

$$\begin{bmatrix} B_u & \nabla h_u \\ \nabla h_u^T & 0 \end{bmatrix}\begin{bmatrix} d_u \\ v_u \end{bmatrix} = -\begin{bmatrix} \nabla \phi_u \\ h_u \end{bmatrix} \quad \nabla h_u = \begin{bmatrix} \nabla_w f_u & 0 \\ \nabla_\zeta f_u & I \end{bmatrix} \quad d = \begin{bmatrix} \Delta w_u \\ \Delta \zeta_u \end{bmatrix} \quad v_u = \begin{bmatrix} \lambda_u \\ \gamma_u \end{bmatrix}$$

Once θ is fixed, can solve for remaining variables independently.

Can exploit each KKT system further.

Structure of Hessian (Gauss-Newton): $B_u = \begin{bmatrix} W_u & 0 \\ 0 & 0 \end{bmatrix}$

38

EVM Decomposition

$$\begin{bmatrix} B_u & \nabla h_u \\ \nabla h_u^T & 0 \end{bmatrix} \begin{bmatrix} d_u \\ v_u \end{bmatrix} = -\begin{bmatrix} \nabla \phi_u \\ h_u \end{bmatrix}$$

<u>Apply Range and Null space Decomposition</u>

- Subspace search directions: $\nabla h_u^T Z_u = 0$, $Y_u^T Z_u = 0$
- Define: $d_u = Z_u p_{Z,u} + Y_u p_{Y,u}$

  Null space step: $p_{Z,u} = -(Z_u^T B_u^{GN} Z_u)^{-1} Z_u^T \nabla \phi_u$

  Range space step: $p_{Y,u} = -(\nabla h_u^T Y_u)^{-1} h_u$
- Note: $p_{Y,u}$ is dependent on $\Delta\theta$, $p_{Z,u}$ is not!

<u>Reconstruct QP Problem is space of $\Delta\theta$</u>

- Sum up contributions from all data sets

$$\underset{\Delta\theta}{Min} \left(\sum_{u=1}^{r} \alpha_u \right)^T \Delta\theta + 1/2\, \Delta\theta^T \left(\sum_{u=1}^{r} \beta_u \right) \Delta\theta$$

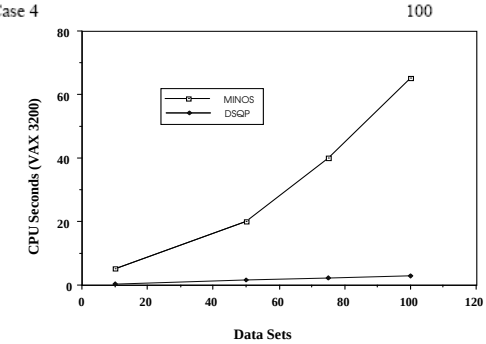
$$s.t. \quad a\text{-}\theta^k \le \Delta\theta \le b\text{-}\theta^k$$

where α_u and β_u are derived from ∇h_u

39

EVM Examples

Example	Model	Number of Data sets	d.o.f
1	$y = \theta_1 + \theta_2 x$	10	12
2	$y = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3$	10	14
3	$\exp(-\theta_1 t \exp(-\theta_2/T)) - y = 0.$	15	32
4	$x_3 = \dfrac{\theta_1 \theta_2^2 \theta_3 x_1 x_2^3}{(1+\theta_1 x_1 + \theta_2 x_2)^3}$	28	59
5, Case 1	$x_2 = \theta_1 + \dfrac{1}{x_1 - \theta_2}$	25	27
5, Case 2		50	52
5, Case 3		75	77
5, Case 4		100	102



40

Full-space NLP Formulation for Parameter Estimation

Original Formulation

$$\min_{x \in \Re^n} f(x)$$
$$\text{s.t} \quad c(x) = 0$$
$$x \geq 0$$

Can generalize for

$$a \leq x \leq b$$

Barrier Approach

$$\min_{x \in \Re^n} \varphi_\mu(x) = f(x) - \mu \sum_{i=1}^{n} \ln x_i$$
$$\text{s.t} \quad c(x) = 0$$

$\Rightarrow$As $\mu \rightarrow 0$, $x^*(\mu) \rightarrow x^*$

Fiacco and McCormick (1968)

41

Solution of the Barrier Problem - IPOPT

$\Rightarrow$**Newton Directions (KKT System)**

$$\begin{aligned} \nabla f(x) + A(x)\lambda - v &= 0 \\ XVe - \mu e &= 0 \\ c(x) &= 0 \end{aligned}$$

$\Rightarrow$ **Reducing the System**

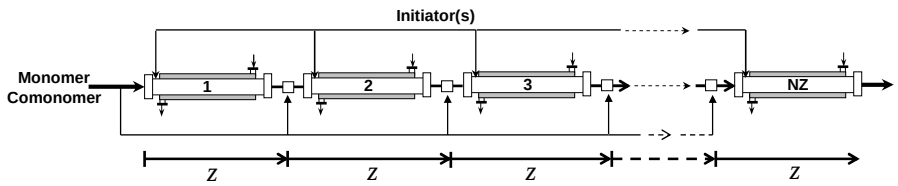
$$d_v = \mu X^{-1} e - v - X^{-1} V d_x$$

$$\begin{bmatrix} Q+\Sigma & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} d_x \\ \lambda^+ \end{bmatrix} = -\begin{bmatrix} \nabla \varphi_\mu \\ c \end{bmatrix} \qquad \Sigma = X^{-1} V$$

What are the Benefits for Parameter Estimation?

42

Large-Scale Parameter Estimation



Material & Energy

Physical Properties

Zone Transitions

$$\left.\begin{cases} \mathbf{F}_j\left[\dfrac{dy_j(z)}{dz}, y_j(z), w_j(z), z, \pi_j, \Pi\right] = 0 \\ \mathbf{G}_j\left[y_j(z), w_j(z), z, \pi_j, \Pi\right] = 0 \\ y_j(0) = \phi(y_{j-1}(z_{L_{j-1}}), F_{f_j}) \end{cases}\right\} j \in \{1..NZ\}$$

500 ODEs
1000 AEs

× **Stiffness + Highly Nonlinear + Parametric Sensitivity + Algebraic Coupling**

Large-Scale Parameter Estimation

- **Complex Kinetic Mechanisms**

Initiator decomposition

$$I_i \xrightarrow{k_{d_i}} 2R \quad i = 1, N_I$$

Chain Initiation

$$R + M_1 \xrightarrow{k_{I_1}} P_{1,0}$$
$$R + M_2 \xrightarrow{k_{I_2}} Q_{0,1}$$

Chain Propagation

$$P_{r,s} + M_1 \xrightarrow{k_{p11}} P_{r+1,s}$$
$$P_{r,s} + M_2 \xrightarrow{k_{p12}} Q_{r,s+1}$$
$$Q_{r,s} + M_1 \xrightarrow{k_{p21}} P_{r+1,s}$$
$$Q_{r,s} + M_2 \xrightarrow{k_{p22}} Q_{r,s+1}$$

Chain Transfer to Monomer

$$P_{r,s} + M_1 \xrightarrow{k_{fm11}} P_{1,0} + M_{r,s}$$
$$P_{r,s} + M_2 \xrightarrow{k_{fm12}} Q_{0,1} + M_{r,s}$$
$$Q_{r,s} + M_1 \xrightarrow{k_{fm21}} P_{1,0} + M_{r,s}$$
$$Q_{r,s} + M_2 \xrightarrow{k_{fm22}} Q_{0,1} + M_{r,s}$$

Chain Transfer to Solvent

$$P_{r,s} + S_i \xrightarrow{k_{si1}} P_{1,0} + M_{r,s}$$
$$Q_{r,s} + S_i \xrightarrow{k_{si2}} Q_{0,1} + M_{r,s}$$

Chain Transfer to Polymer

$$P_{r,s} + M_{x,y} \xrightarrow{k_{fp11}} P_{x,y} + M_{r,s}$$
$$P_{r,s} + M_{x,y} \xrightarrow{k_{fp12}} Q_{x,y} + M_{r,s}$$
$$Q_{r,s} + M_{x,y} \xrightarrow{k_{fp21}} P_{x,y} + M_{r,s}$$
$$Q_{r,s} + M_{x,y} \xrightarrow{k_{fp22}} Q_{x,y} + M_{r,s}$$

Termination by Combination

$$P_{r,s} + P_{x,y} \xrightarrow{k_{tc11}} M_{r+x,s+y}$$
$$P_{r,s} + Q_{x,y} \xrightarrow{k_{tc12}} M_{r+x,s+y}$$
$$Q_{r,s} + Q_{x,y} \xrightarrow{k_{tc22}} M_{r+x,s+y}$$

Termination by Disproportionation

$$P_{r,s} + P_{x,y} \xrightarrow{k_{td11}} M_{r,s} + M_{x,y}$$
$$P_{r,s} + Q_{x,y} \xrightarrow{k_{td12}} M_{r,s} + M_{x,y}$$
$$Q_{r,s} + Q_{x,y} \xrightarrow{k_{td22}} M_{r,s} + M_{x,y}$$

Backbitting

$$P_{r,s} \xrightarrow{k_{b1}} P_{r,s} \ or \ Q_{r,s}$$
$$P_{r,s} \xrightarrow{k_{b2}} Q_{r,s} \ or \ P_{r,s}$$

β-scission

$$P_{r,s} \xrightarrow{k_{\beta 1}} M^{=}_{r,s} + P_{1,0}$$
$$P_{r,s} \xrightarrow{k_{\beta 2}} M^{=}_{r,s} + Q_{0,1}$$

$$k = k_0 \, exp\left(-\frac{E_a + P\,E_v}{RT}\right)$$

~ 35 Elementary Reactions
~100 Kinetic Parameters

Large-Scale Parameter Estimation

- **Parameter Estimation for Industrial Applications**
  - **Use Rigorous Model to Match Plant Data Directly**
  - **Start with Standard Least-Squares Formulation**

$$\min_{\Pi,\pi_{k,j}} \sum_{k=1}^{NS}\sum_{j=1}^{NZ}\sum_{i=1}^{NM(j)} \left(y_{k,j}(z_i)-y^M_{k,j,i}\right)^T \mathbf{V_y^{-1}} \left(y_{k,j}(z_i)-y^M_{k,j,i}\right) + \sum_{k=1}^{NS}\left(w_{k,NZ}-w^M_{k,NZ}\right)^T \mathbf{V_w^{-1}}\left(w_{k,NZ}-w^M_{k,NZ}\right)$$

Least-Squares

$$s.t.$$

$$\mathbf{F}_{k,j}\left[\frac{dy_{k,j}(z)}{dz}, y_{k,j}(z), w_{k,j}(z), z, \pi_{k,j}, \Pi\right]=0$$

$$\mathbf{G}_{k,j}\left[y_{k,j}(z), w_{k,j}(z), z, \pi_{k,j}, \Pi\right] 0$$

$$y_{k,j}(0)=\phi(y_{k,j-1}(z_{L_{k,j-1}}), F_{f_{k,j}})$$

$$j\in\{1..NZ\},\quad k\in\{1..NS\}$$

Rigorous Reactor Model

- **Special Case of Multi-Stage Dynamic Optimization Problem**
  - **Solve using Simultaneous Collocation-Based Approach**

1 data set		6 data sets
500 ODEs	x 6 →	3000 ODEs
1000 AEs		6000 AEs

45

Large-Scale Parameter Estimation

- **Multi-Zone Tubular Reactor –** Quasi Steady-State
  - **Data Sets:** Operating Conditions and Properties for Different Grades
  - **Match:** Temperature Profiles and Product Properties
    - **On-line Adjusting Parameters** → Track Evolution of Disturbances
    - **Kinetic Parameters** → Development and Discrimination among Rigorous Models
- **Results**
  - **Single Data Set** (On-line Adjusting Parameters)

Grade	Constraints	Parameters	LB	UB	Iterations	CPUs	NZJ	NZH
A	11955	32	374	361	11	17.03	166425	87954
B	11283	32	374	361	8	10.06	138666	76890

  - **Multiple Data Sets** (On-line Adjusting Parameters + Kinetics)

Data Sets	Constraints	DOF	LB	UB	Iterations	CPUs	NZJ	NZH
3	33900	121	1246	1207	68	451.51	520275	552738
6	68421	217	2467	2389	58	900.21	1058412	1119258

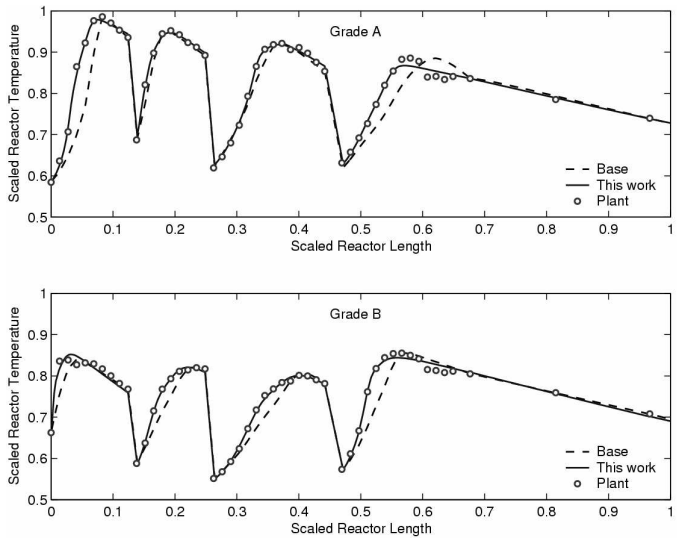
Bottleneck (Memory Requirements)
Factorization Step

$$\begin{bmatrix} W(x_k,\lambda_k) & A(x_k) & -I \\ A(x_k)^T & 0 & 0 \\ V_k & 0 & X_k \end{bmatrix}\begin{bmatrix}\Delta x\\ \Delta\lambda\\ \Delta\nu\end{bmatrix} = -\begin{bmatrix}\nabla f(x_k)+A(x_k)\lambda_k-\nu_k \\ c(x_k) \\ X_kV_ke-\mu_\ell e\end{bmatrix}$$

46

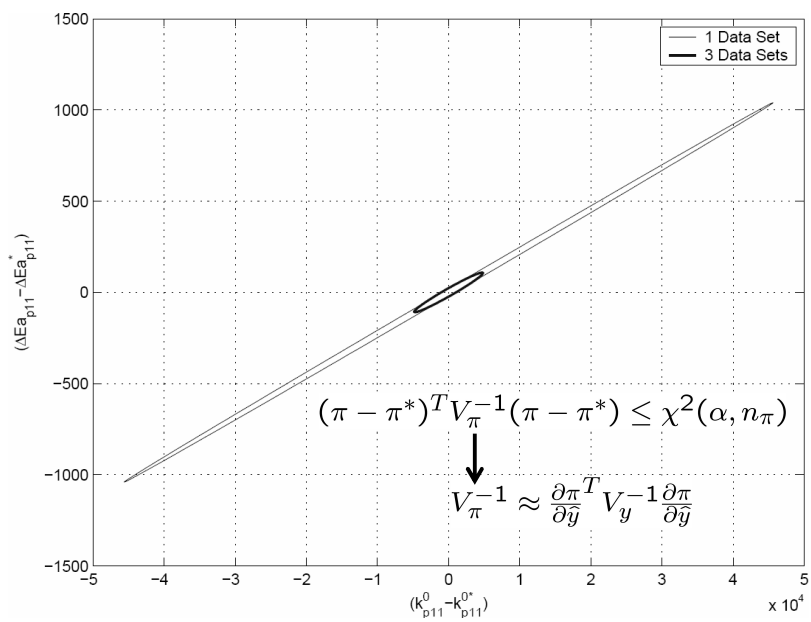
Large-Scale Parameter Estimation

Improved Match of Reactor Temperatures Profile



47

IPOPT Factorization Byproduct



$$(\pi - \pi^*)^T V_\pi^{-1} (\pi - \pi^*) \leq \chi^2(\alpha, n_\pi)$$

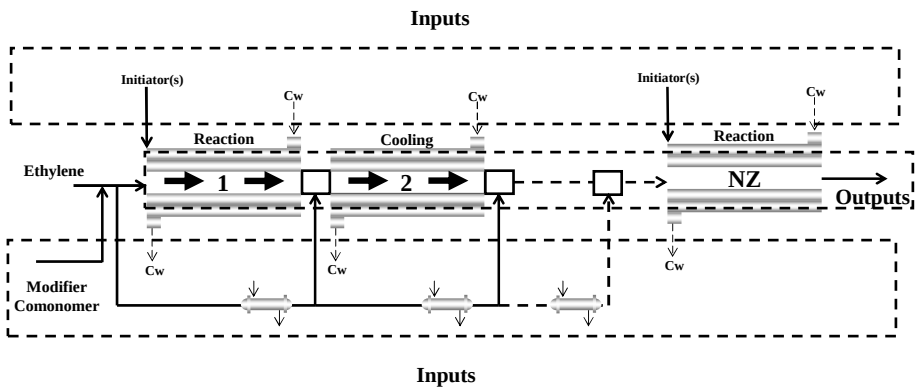
$$V_\pi^{-1} \approx \frac{\partial \pi}{\partial \hat{y}}^T V_y^{-1} \frac{\partial \pi}{\partial \hat{y}}$$

48

Advanced Regression Methods

Errors-In-Variables (EVM)

- **Standard Least Squares - Errors in Output Variables - *Biased* Parameters**
- **EVM - Errors in Output *AND* Input Variables - *Unbiased* Parameters**



Advanced Regression Methods

Errors-In-Variables (EVM)

EVM Drawback - Degrees of Freedom $DOF = \Pi + \sum_{k-1}^{NS}\sum_{j=1}^{NZ}\pi_{k,j} + \sum_{k=1}^{NS}\sum_{j=1}^{NZ}\sum_{i=1}^{NM_u(j)} u_{k,j,i}$

Formulation is Straightforward

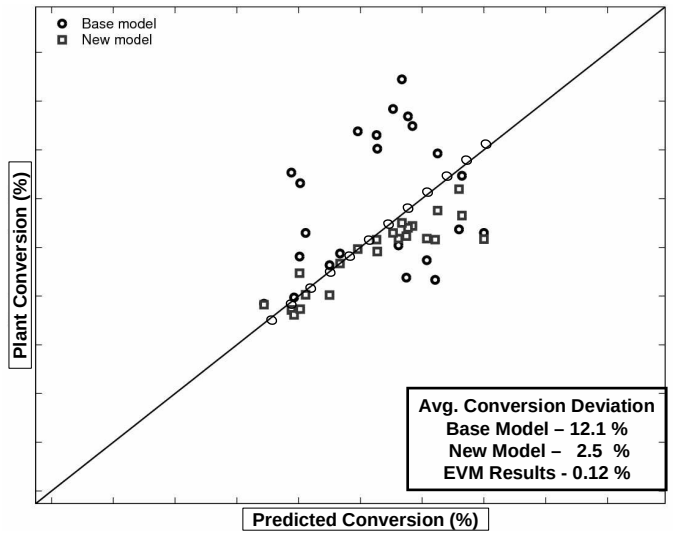
$$\min_{\Pi,\,\pi_{k,j},\,u_{k,j}} \sum_{k=1}^{NS}\sum_{j=1}^{NZ}\sum_{i=1}^{NM(j)} \left(y_{k,j}(z_i) - y^M_{k,j,i}\right)^T \mathbf{V_y^{-1}} \left(y_{k,j}(z_i) - y^M_{k,j,i}\right)$$
$$+ \sum_{k=1}^{NS} \left(w_{k,NZ} - w^M_{k,NZ}\right)^T \mathbf{V_w^{-1}} \left(w_{k,NZ} - w^M_{k,NZ}\right)$$
$$+ \sum_{k=1}^{NS}\sum_{j=1}^{NZ} \left(u_{k,j} - u^M_{k,j}\right)^T \mathbf{V_u^{-1}} \left(u_{k,j} - u^M_{k,j}\right)$$

EVM vs. Standard Least Squares

Data Sets	Constraints	DOF	LB	UB	Iterations	CPUs	NZJ	NZH
6 (EVM)	68627	529	2653	2575	71	1010.74	1059512	1119780
6 (SLS)	68421	217	2467	2389	58	900.21	1058412	1119258

Industrial Case Study

- **Results - Reactor Overall Monomer Conversion (up to 20 Different Grades)**

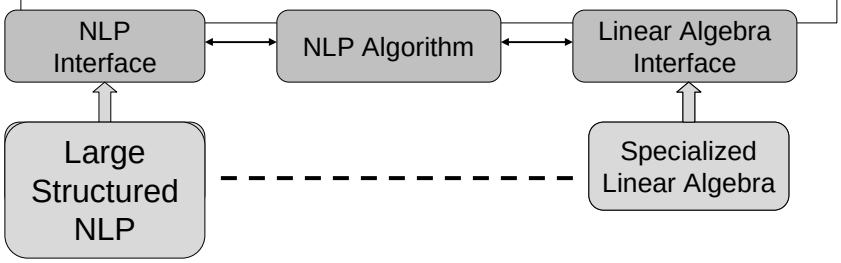


Carnegie Mellon 51

Next Generation IPOPT

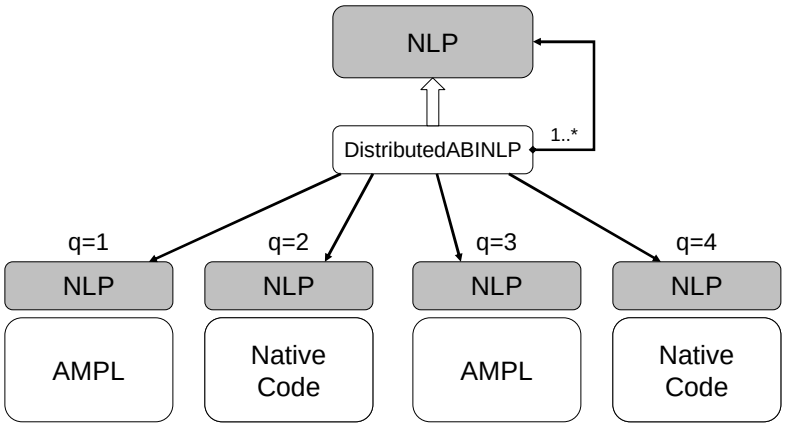
- IPOPT 3.2, Fall, 2006, CPL (www.coin-or.org)
  - Based on Fortran version
  - Object-oriented, NLP Solver
    - Primal-Dual Interior Point method
    - Full space - exact Hessian information
    - Monotone/Adaptive μ update
    - Filter line search strategy
  - Flexible algorithm structure
  - Interfacing to other linear solvers
  - Modeling Environments - AMPL, AIMMS, MATLAB…
- *Ideal for Internal Decomposition*
  - Consistent linear system structure at every iteration
  - Separation of algorithm and specialized linear algebra

High Level IPOPT Design



- Provide structured NLP elements
  - Hessian, jacobian, gradients, residuals, variables,
- Vector operations
  - BLAS, norms, dot products, axpy, element- wise, max, min, etc.
- Matrix-Vector operations
  - $M\,v$, $M^T v$
- Solution of Linear KKT system

Modeling the Block NLP



54

Parameter Estimation in Parallel Architectures

Exploit Structure of KKT Matrix – *Laird, Biegler 2006*

$$\min_{\Pi,\pi_{k,j}} \sum_{k=1}^{NS}\sum_{j=1}^{NZ}\sum_{i=1}^{NM(j)} \left(y_{k,j}(z_i)-y^M_{k,j,i}\right)^T \mathbf{V}_{\mathbf{y}}^{-1}\left(y_{k,j}(z_i)-y^M_{k,j,i}\right) + \sum_{k=1}^{NS}\left(w_{k,NZ}-w^M_{k,NZ}\right)^T \mathbf{V}_{\mathbf{w}}^{-1}\left(w_{k,NZ}-w^M_{k,NZ}\right)$$

$$s.t.\quad \mathbf{F}_{k,j}\left[\frac{dy_{k,j}(z)}{dz}, y_{k,j}(z), w_{k,j}(z), z, \pi_{k,j}, \Pi\right]=0$$
$$\mathbf{G}_{k,j}\left[y_{k,j}(z), w_{k,j}(z), z, \pi_{k,j}, \Pi\right]=0$$
$$y_{k,j}(0)=\phi(y_{k,j-1}(z_{L_{k,j-1}}), F_{f_{k,j}})$$
$$j\in\{1..NZ\},\quad k\in\{1..NS\}$$

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & c(x)=0 \\ & x\geq 0 \end{aligned}$$

$$\begin{aligned} \min \quad & \sum_{k=1}^{N_S} f_k(x_k) \\ \text{s.t.} \quad & c_k(x_k,\Pi)=0 \\ & x_k,\Pi\geq 0 \end{aligned}$$

$$\begin{bmatrix} H_k & A_k & -I \\ A_k^T & 0 & 0 \\ V_k & 0 & X_k \end{bmatrix}\begin{bmatrix} \Delta x \\ \Delta\lambda \\ \Delta\nu \end{bmatrix} = -\begin{bmatrix} r_x \\ r_\lambda \\ X_kV_ke-\mu_\ell e \end{bmatrix}$$

Direct Factorization MA27
Memory Bottlenecks
Factorization Time *Scales Superlinearly* with Data sets

$$\begin{bmatrix} \mathbf{K}_1 & & & & Q_1 \\ & \mathbf{K}_2 & & & Q_2 \\ & & \ddots & & \vdots \\ & & & \mathbf{K}_N & Q_N \\ Q_1^T & Q_2^T & \dots & Q_N^T & D_\Pi \end{bmatrix}\begin{bmatrix} \Delta s_1 \\ \Delta s_2 \\ \vdots \\ \Delta s_N \\ \Delta\Pi \end{bmatrix} = -\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \\ r_\Pi \end{bmatrix}$$

Block-bordered Diagonal Structure
Coarse-Grained Parallelization

Parameter Estimation in Parallel Architectures

Sch

$$\begin{bmatrix} \Delta s_1 \\ \Delta s_2 \\ \vdots \\ \Delta s_N \\ \Delta\Pi \end{bmatrix} = -\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \\ r_\Pi \end{bmatrix}$$

1) **Parallelizable**
Sparse Factorization $\mathbf{K}_i$ in each Block
Inertia Correction

2) **Non-parallelizable**
Dense Factorization of $\mathbf{S} \rightarrow dim\{\Pi\}$

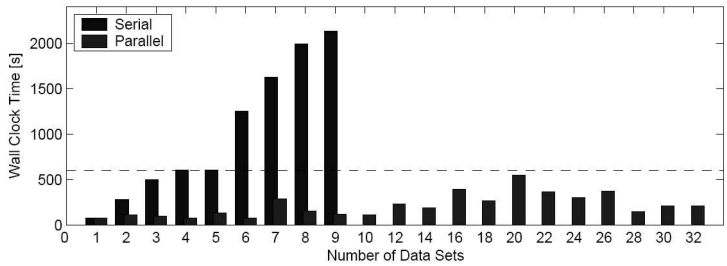
3) **Parallelizable**
Final Backsolve

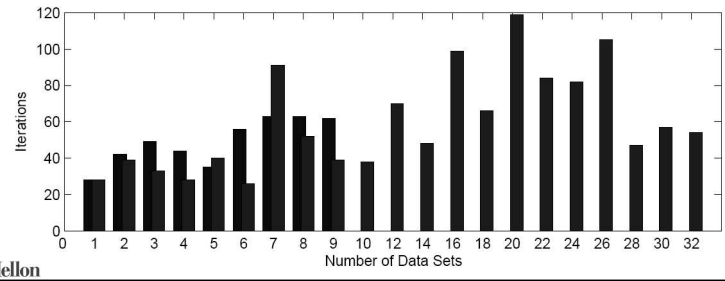
Parameter Estimation in Parallel Architectures

Computational Results – LDPE Reactor EVM Problem

Beowulf Cluster – ChemE Department, Carnegie Mellon University

Intel Pentium 4 Xeon 2.4ghz, 2gb RAM



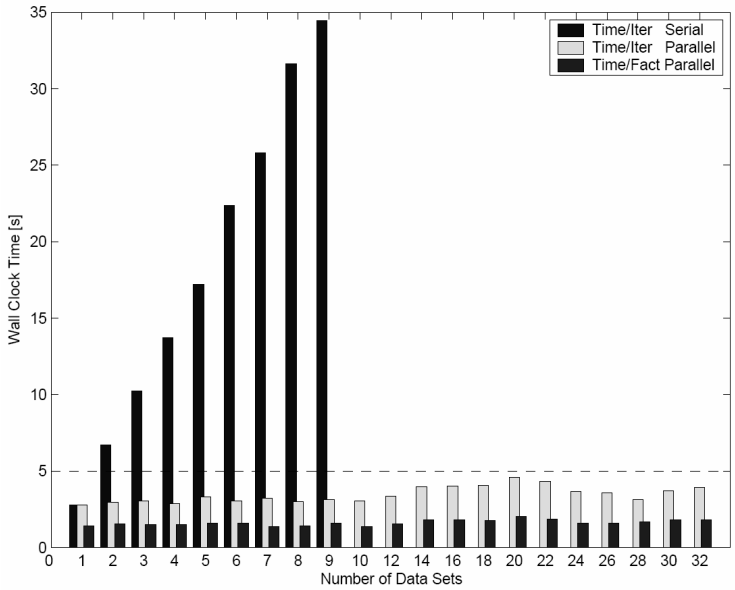


Carnegie Mellon

57

Parameter Estimation in Parallel Architectures

Computational Results – LDPE Reactor EVM Problem



Carnegie Mellon

58

Conclusions – Parameter Estimation

- Trust Region (Levenberg-Marquardt) methods: standard for unconstrained problems - MINPACK, NaG, NL2SOL, Harwell
- Constrained problem formulations - more model flexibility
- SQP codes adapted to exploit least squares structure, faster methods
- EVM problems - expensive for conventional optimization codes
  - Many degrees of freedom for optimization
  - Decomposition of KKT conditions required
  - ODRPACK (netlib) developed for $y_u = f_u(\theta)$
- Large-scale SQP methods developed for:
  - Parameter estimation
  - EVM methods
  - Data Reconciliation
- IPOPT has useful characteristics for large-scale parameter estimation

59

Optimization Algorithms for Data Reconciliation

Introduction to Data Reconciliation

M-Estimators
- Bayesian Forms
- Fair Function
- Redescending Estimator

Akaike Information Criterion

Mixed Integer Formulations

Static Examples

Dynamic Examples

Conclusions

60

Introduction

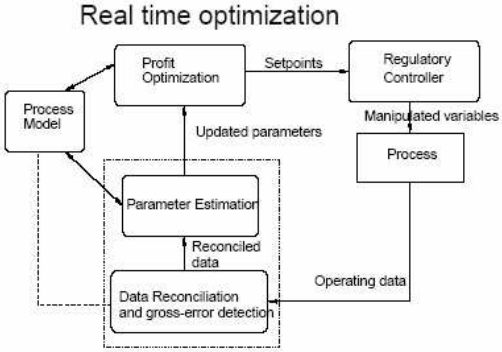
- Data contaminated by errors
  - Random errors:
    - High frequency noise
    - Typically $\sim \mathcal{N}(0, \Sigma^2)$
  - Gross errors and outliers
    - Systematic occurrences: biases, drifts
    - Not modeled by distributions
- Contaminated data can bias estimates
- Difficult to converge NLP when data noisy
- Inputs may be contaminated

61

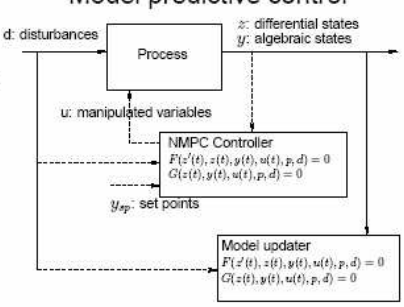
Data Reconciliation Framework

- Industrial applications of data reconciliation and parameter estimation (DRPE)

Real time optimization

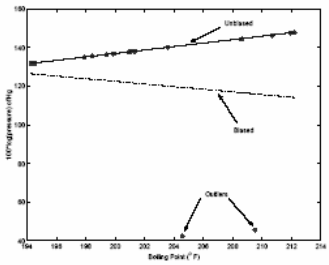


Model predictive control



Effects of Gross Error in Regression

Example: Boiling point of H_2O vs. atm. pressure with 2 outliers (Weisberg, 1985)



Sources of gross errors: Broken gauges, process leaks, improperly used measurement devices, operator errors

Identify and negate effects of gross errors

63

Data Reconciliation – Literature Review

- Sequential χ^2 tests:
  - Crowe et al. (1983), Madron (1985), Kao et al. (1992)
- Combinatorial
  - Serth and Heenan (1986), Narasimhan and Mah (1987): Serial Combination and Generalized Likelihood Ratio tests
  - Crowe (1989): Max. power test
  - Rollins et al. (1996): Linear Combination Test based on χ^2 tests
  - Bagajewicz et al. (1998): Graph theory based, linear ODEs
- Simultaneous
  - Tjoa and Biegler (1991): Contaminated normal - difficult to tune, cannot detect biases
  - Albuquerque and Biegler (1996): Fair function - easy to use, no explicit outlier identification

64

Treatment of Outliers: M-Estimators

M-Estimator: modeling the influence of residual outliers through modification of maximum likelihood (ML) functions

Bayesian Statistics and Bernoulli Trials

- Statistical definitions and ML function
- Allows statistical inference

Robust Statistics

- Modification without inferential aspects

Akaike Information Criterion (AIC)

- Based on ML extended to discrete parameters

65

Bayesian Approach: Bernoulli Trial

Assume separate probability distributions for random and gross errors:

$$P(\varepsilon_i \mid \theta, X_i) = R(\varepsilon_i \mid \theta, X_i)\Pi(R) + G(\varepsilon_i \mid \theta, X_i)\Pi(G)$$

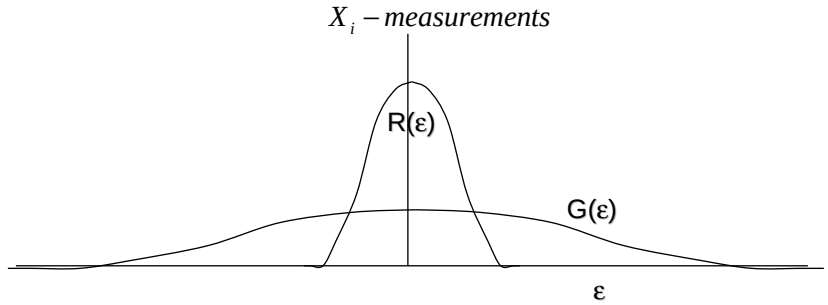
$$P, R, G - \text{ probability distributions}$$

$$\Pi(R), \Pi(G) - \text{ prior distributions}$$

$$\varepsilon_i \text{- measurement error}$$

$$\theta - \text{ parameters}$$

$$X_i - \text{measurements}$$



66

Bayesian Approach: Bernoulli Trial

Combine both distributions into ML function:

$$P(\varepsilon_i \mid \theta, X_i) = R(\varepsilon_i \mid \theta, X_i)\Pi(R) + G(\varepsilon_i \mid \theta, X_i)\Pi(G)$$

$$P(\varepsilon_i \mid \theta, X_i) = R(\varepsilon_i \mid \theta, X_i)(1-p) + G(\varepsilon_i \mid \theta, X_i)\,p$$

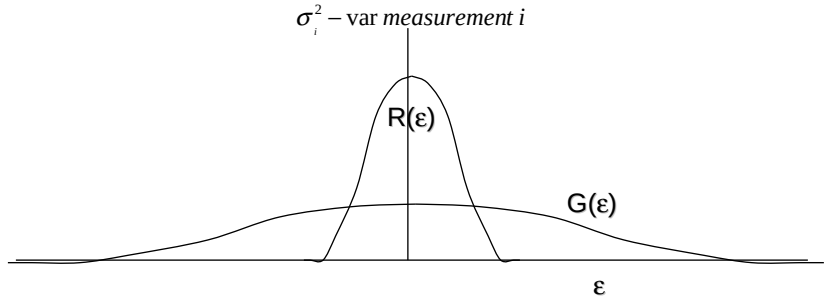
$$P(\varepsilon_i \mid \theta, X_i) = \exp(-\varepsilon_i^2/\sigma_i^2)(1-p)/\sigma_i + \exp(-\varepsilon_i^2/(b\sigma_i)^2)\,p/(\sigma_i b)$$

$$\max \sum_i \log(P(\varepsilon_i \mid \theta, X_i))$$

$$p \text{ - } fraction\ gross\ errors$$

$$b - \ scale\ of\ distributions$$

$$\sigma_i^2 - \text{var}\ measurement\ i$$

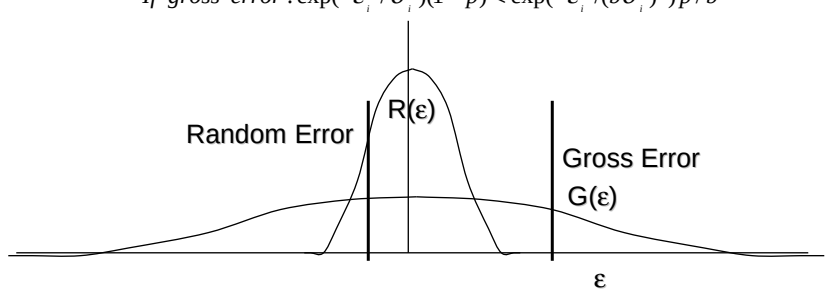


Bayesian Approach: Gross Error Test

Solve optimization problem and perform gross error tests at solution:

$$\max \sum_i \log(\exp(-\varepsilon_i^2/\sigma_i^2)(1-p) + \exp(-\varepsilon_i^2/(b\sigma_i)^2)\,p/b)$$

$$If\ \ gross\ \ error: \exp(-\varepsilon_i^2/\sigma_i^2)(1-p) < \exp(-\varepsilon_i^2/(b\sigma_i)^2)\,p/b$$



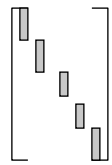
+ Statistical basis for determining gross errors
-Assumes gross errors follow proposed distribution
-Not robust to deviations from assumptions

Example (Pai and Fisher, 1988)

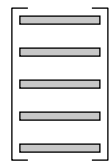
$$0.5(x_1)^2 - 0.7x_2 + x_3u_1 + (x_2)^2u_1u_2 + 2x_3(u_3)^2 - 255.8 = 0$$
$$x_1 - 2x_2 + 3x_1x_2 - 2x_2u_1 - x_2u_2u_3 + 111.2 = 0$$
$$x_3u_1 - x_1 + 3x_1x_3 - 2x_2u_1 - x_2u_2u_3 + 111.2 = 0$$
$$x_4 - x_1 - (x_3)^2 + u_2 + 3u_3 = 0$$
$$x_5 - 2x_3u_2u_3 = 0$$
$$2x_1 + x_2x_3u_1 + u_2 - u_3 - 126.6 = 0$$

5 measured variables (x)
3 unmeasured variables (u)
b= 20, p=0.2, σ=0.1

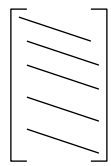
Introduce 20% gross error into
500 data sets generated randomly



100 right
0 wrong



87 right
0 wrong



99 right
1 wrong

69

Robust Statistics for Gross Errors

- Common Assumption: Data with normally distributed random errors ⇒ likelihood function is *least squares estimator*
- likelihood of data corrupted with outliers: difficult to determine
- Estimators derived from fixed probability distribution: not justified
- Robust Estimators:
  - largely distribution independent
  - unbiased estimates when data from ideal distribution
  - Insensitive to deviations from ideality
  - Add less weight to outlying measurements ⇒ *protect* other measurements from being corrupted

70

Robust Statistics Properties

Find estimators that are insensitive to deviations in assumptions of noise distribution.

F – assumed distribution of data

G – actual distribution of data

$$\delta(F,G) < \varepsilon \Rightarrow \delta(T(F),T(G)) < \gamma$$

T(-) – test statistic or distribution of estimator

<u>Example:</u>

S1: [2.0, 2.1, 2.2, 2.3, 2.4] Mean = 2.2, Median=2.2

S2: [2.0, 2.1, 2.2, 2.3, 24] Mean = 6.56, Median=2.2

➔ Median is a robust statistic, Mean is not.

71

Huber Class of Robust Estimators (1981)

Fair Function (Rey, 1988)

$$\rho_{Fi} = C^2\left[\frac{|\varepsilon_i|}{C} - \log\left\{1+\frac{|\varepsilon_i|}{C}\right\}\right]$$

- Use $\sum_i \rho_{Fi}$ instead of least squares objective
- Convex function of <u>normalized</u> residual
- Small residuals – quadratic behavior
- Large residuals – linear behavior
- C – Tuning parameter based on Cramer-Rao bound (trade off efficiency (high C) with robustness (low C))

$$C = 0.21529\left(\frac{\varphi - 0.63662}{1-\varphi}\right) \quad where \quad \varphi = \frac{\sigma^2}{V_{Fair}}$$

72

Hampel Class of Robust Estimators

Three Part Redescending Estimator (Hampel, 1974)

$$\rho_{H_i} = \begin{cases} \frac{1}{2}\varepsilon_i^2, & 0 \le |\varepsilon_i| \le a \\ a\,|\varepsilon_i| - \frac{a^2}{2}, & a < |\varepsilon_i| \le b \\ ab - \frac{a^2}{2} + (c-b)\frac{a}{2}\left[1 - \left(\frac{c-|\varepsilon_i|}{c-b}\right)^2\right], & b < |\varepsilon_i| \le c \\ ab - \frac{a^2}{2} + (c-b)\frac{a}{2}, & |\varepsilon_i| > c, \end{cases}$$

Three parameters define regions: c > b+2a

Quadratic, linear and constant parts involving normalized residual

Nonsmooth – requires some smoothing approximations

Not clear how to tune a, b, c

73

M-Estimators in Robust Statistics

M-Estimators

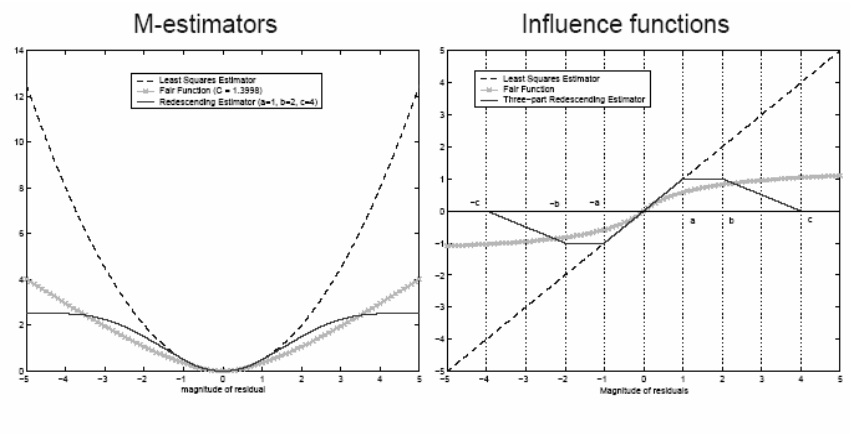
- Overall likelihood of N observations: $\mathcal{L} = \prod_{i=1}^{N} \ell(x_i|p)$
- M-estimator = $\rho^M = -log\ \mathcal{L} = -\sum_{i=1}^{N} log\ \ell(x_i|p)$

M-Estimators Used $\mathcal{IF} = \psi(\varepsilon_0) = \lim_{t\to 0} \frac{T[(1-t)f + t\delta(\varepsilon - \varepsilon_0)] - T[f]}{t}$

1. Least Squares Estimator (Non robust)
   - Severely biased by outliers
2. Fair Function (Robust)
   - $\mathcal{IF}$ is bounded as $\varepsilon \to \infty$
3. Three part Redescending Estimator (Robust)
   - $\mathcal{IF} \to 0$ as $\varepsilon \to \infty$

74

Properties of M-Estimators



75

Properties of M-Estimators

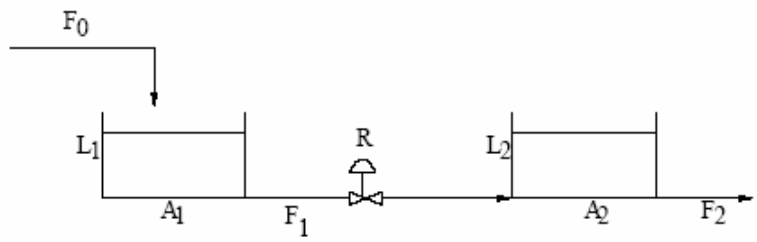
	Bayesian	Robust
Distributional Assumptions?	Yes	No
Sensitive to Assump. Deviations?	Yes	No
Statistical Inference at Solution?	Yes	No
Incorporate prior knowledge?	Yes	No

-For robust case, apply exploratory statistics (e.g., boxplots) at solution

- Need data redundancy to identify gross errors

76

Dynamic Data Reconciliation

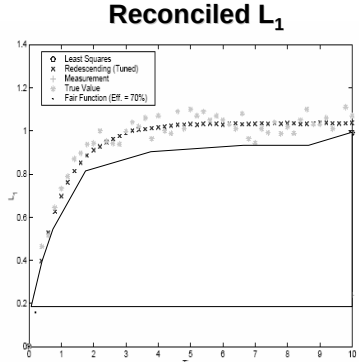


- Measurements: Flows F_0, F_1, F_2 & heights L_1 and L_2
- DAE system
- Parameters: Tank areas: $1/A_1$ & $1/A_2$

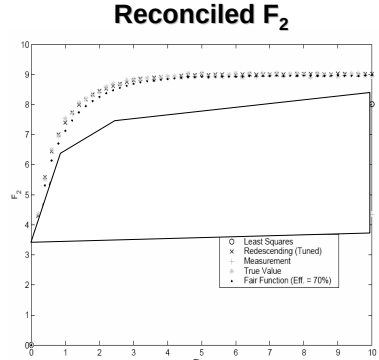
77

Case 1: Comparison of Results

Reconciled L_1



Reconciled F_2



•Discretize DAEs and solve as large scale NLP problem with appropriate objective function

•Add random outlier noise to flow and level data

78

Tank Example – Case 1

- Outliers drawn from broad random distribution (distributional assumption satisfied)
- Data reconciliation and parameter estimates done well by
- Bayesian approach and M-estimators

Parameters	$1/A_1$	$1/A_2$
True Values	0.5	0.5
Least Squares	0.698	0.503
Bayesian	0.490	0.50
Fair Function	0.501	0.501
Redesc. Tuned	0.50	0.50

79

Tank Example – Case 2

- Outliers systematic (distributional assumption violated); measurements for L_1 and F_2 stuck
- Data reconciliation and parameter estimation poor with Bayesian approch
- Data reconciliation does and parameter estimates done better by
- M-estimators

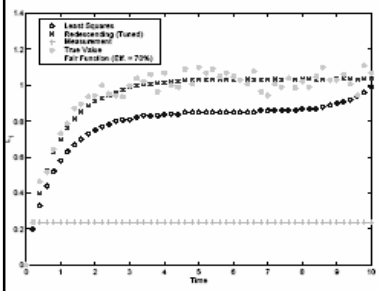
Parameters	$1/A_1$	$1/A_2$
True Values	0.5	0.5
Least Squares	0.25*	0.55
Bayesian	0.25	0.25
Fair Function	0.25	0.439
Redesc. Tuned	0.499	0.500

*lower bound

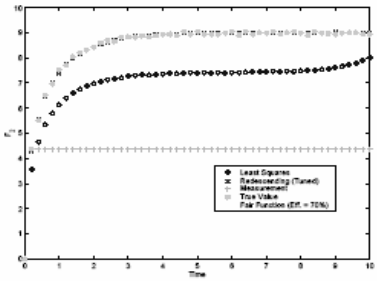
80

Case 2: Comparison of Results

Reconciled L_1



Reconciled F_2



- Least squares: Very poor DRPE
- Fair function: Better than least squares but inadequate
- Redescending: Avoids bounds on $1/A_1$ and $1/A_2$. Tuned for best result

Redescending Estimators superior for problems with largely corrupted data

Gross Errors Modeled with AIC

Common statistical framework for combinatorial and robust DRPE strategies?

- DR & Gross error detection $\equiv$ model discrimination and PE
- Data models $\Rightarrow$ partitions of residuals into random and gross errors
- Parameters = # of outliers + model parameters
- Obtain most likely model and its parameters
- Max. likelihood estimators (MLE): asymptotically efficient $\Rightarrow$ likelihood function = sensitive criterion of deviation of parameters from true values
- Akaike Information Criterion ($\mathcal{AIC}$): Distance between fitted and true models

$$\mathcal{AIC} = -2 \cdot \log(\text{maximum likelihood}) + 2(\text{\# independently adjusted parameters in model})$$

82

AIC for Data Reconciliation

Objective Function

$$\mathcal{AIC} = -2log(\text{maximum likelihood}) + 2(\#\text{ independently adjusted parameters within the model}) .$$

$$\mathcal{AIC} = E(S) = 2\sum_{i=1}^{N} -log(\ell(\varepsilon(i,p),i,p)) + 2\ dim(p),$$

- **Contains discrete and continuous variables**
- **dim(p) = # gross errors in problem**
- **Leads to MINLP problem**
- **Can be simplified to MILP for linear systems**
- **AIC can also be used as an off-line objective for calibration**

83

MINLP formulation

(Yamamura et al, 1988; Arora and B, 2001)

$$\min_{\substack{x_i,\mu_i,\\ y_i,z_i}} \sum_{i=1}^{n} \left[(x_i^M - x_i)/\sigma_i - \mu_i/\sigma_i\right]^2 + 2\cdot\sum_{i=1}^{n} y_i$$

$$\text{s.t. } \boxed{Ax = 0}$$

$$\mu_i \le U_i y_i$$

$$-\mu_i \le U_i y_i$$

$$\mu_i - z_i U_i - z_i L_i + L_i y_i \le 0$$

$$-\mu_i + z_i U_i + z_i L_i + L_i y_i \le L_i + U_i$$

$$x_i \ge 0$$

$$z_i \le y_i$$

$$y_i, z_i \in \{0,1\}.$$

-Direct minimization of AIC function

-Model is for linear system, but straightforward extension for nonlinear systems

- μ_i = positive or negative biases, $U \ge |\mu_i| \ge L$

-Binary variables enforce this:

y = 0,1: existence of gross error

z = 0,1: positive or negative sign

84

MILP Simplification
(Soderstrom et al., 2000)

$$\min_{\substack{x_i,\mu_i,\\ y_i,z_i}} \sum_{i=1}^{n} \left|(x_i^M - x_i)/\sigma_i - \mu_i/\sigma_i\right| + \sum_{i=1}^{n} w_i \cdot y_i$$
$$\begin{aligned}
\text{s.t.}\quad & Ax = 0\\
& \mu_i \le U_i y_i\\
& -\mu_i \le U_i y_i\\
& \mu_i - z_i U_i - z_i L_i + L_i y_i \le 0\\
& -\mu_i + z_i U_i + z_i L_i + L_i y_i \le L_i + U_i\\
& z_i \le y_i\\
& y_i, z_i \in \{0,1\},
\end{aligned}$$

$$\min_{\substack{x_i,\mu_i,\\ y_i,z_i}} \sum_{i=1}^{n} (r_i + q_i)/\sigma_i + \sum_{i=1}^{n} w_i \cdot y_i$$
$$\begin{aligned}
\text{s.t.}\quad & Ax = 0\\
& x_i^M - (x_i + \mu_i) = r_i - q_i\\
& \mu_i \le U_i y_i\\
& -\mu_i \le U_i y_i\\
& \mu_i - z_i U_i - z_i L_i + L_i y_i \le 0\\
& -\mu_i + z_i U_i + z_i L_i + L_i y_i \le L_i + U_i\\
& z_i \le y_i\\
& r_i, q_i \ge 0\\
& y_i, z_i \in \{0,1\}.
\end{aligned}$$

- Quadratic terms become absolute values (like Huber)
- Weighting values (w_i) not clear
- Limited to linear models
- Much faster solution times

85

Mixed Integer Formulation for Gross Errors

Minimize $\mathcal{AIC}$ by solving *MINLP*:

$$\begin{aligned}
\min_{x_i,\mu_i,y_i} \quad & \sum_{i=1}^{n} \left[\frac{(x_i^M - x_i)}{\sigma_i} - \frac{\mu_i}{\sigma_i}\right]^2 + 2 \cdot \sum_{i=1}^{n} y_i\\
\text{s.t.} \quad & Ax = 0, \qquad (-1)\\
& |\mu_i| \le U_i y_i,\\
& |\mu_i| \ge L_i y_i,\\
& y_i \in \{0,1\},\\
& x_i \ge 0.
\end{aligned}$$

MILP (Soderstrom et al., 2000):

$$\begin{aligned}
\min_{x_i,\mu_i,y_i} \quad & \sum_{i=1}^{n} \left|\frac{(x_i^M - x_i)}{\sigma_i} - \frac{\mu_i}{\sigma_i}\right| + \sum_{i=1}^{n} w_i \cdot y_i\\
\text{s.t.} \quad & Ax = 0, \qquad (-2)\\
& |\mu_i| \le U_i y_i,\\
& |\mu_i| \ge L_i y_i,\\
& y_i \in \{0,1\},\\
& x_i \ge 0.
\end{aligned}$$

w_i to obtain independent biases

- Both MINLP and MILP: Expensive for EVM problems
- In MILP: Choosing weights not entirely clear
- Neither suitable for large EVM problems

Cheaper + reliable DRPE in presence of outliers $\Rightarrow$ use Robust Statistics

86

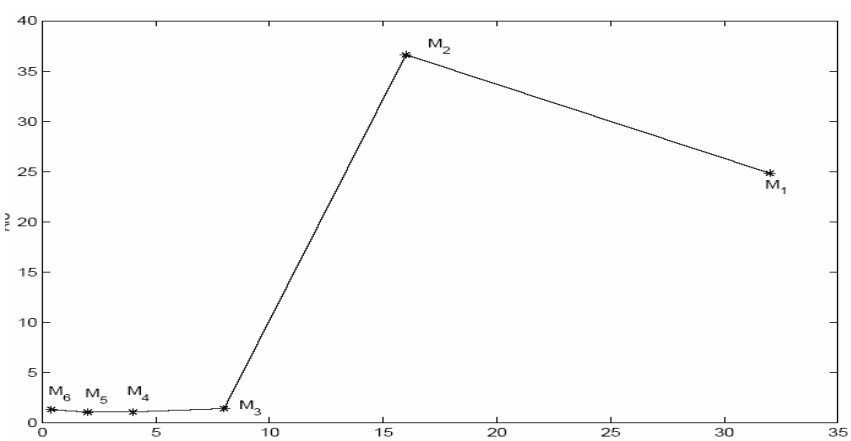
Tuning the Redescending Parameters

- Idea: *Minimize* $\mathcal{AIC}$ associated with the redescending estimator
- Tuning constants related to contamination in data
- Use size of redescending region to assign b and a to obtain c from $c = b + 2a$
- Simple Two-Step Tuning Procedure

1. Select range of redescending estimators 2. Perform DRPE and obtain values of $\mathcal{AIC}$	Step 1
3. Obtain bounds for location of minimum 4. Golden section search for c 5. $b = c/2$, $a = (c - b)/2$	Step 2

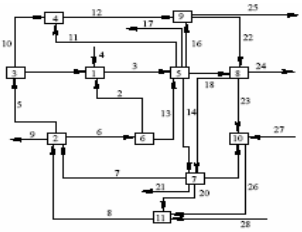
87

Tune Redescending Estimator Using AIC Tanks Example



88

Steam Metering Example



- 28 streams, 11 nodes $\Rightarrow$ 11 linear equalities, 28 variables
- *Moving horizon* of data used $X^M = \begin{bmatrix} x_{11}^M & x_{12}^M & \cdots & x_{1H}^M \\ x_{21}^M & x_{22}^M & \cdots & x_{2H}^M \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^M & x_{n2}^M & \cdots & x_{nH}^M \end{bmatrix}$
- Sizes of horizon attempted: 5, 10, 20, 40
- 3, 5, 7 streams biased; bias propagation
- 100 moving horizons for fixed location of bias
- *Smoothed* redescending estimator for use in *GAMS*

89

Steam Metering Results

$$\mathbf{OP} = \frac{\text{\# biases correctly identified}}{\text{\# biases simulated}}, \quad \mathbf{AVT1} = \frac{\text{\# unbiased wrongly identified}}{\text{\# simulation trials}}.$$

Sample results for horizon size = 5, 3 biases:

Method	$\mathbf{w_i}$	OP	AVT1	$\mathcal{AIC}$	CPU (s)/ horizon
Fair function	–	0.640	3.907	–	–
RE (a, b, c)					
0.5, 1, 2	–	0.673	5.820	3.491	1.114
1, 2, 4	–	0.626	2.065	14.112	1.075
2, 4, 8	–	0.571	0.305	40.299	1.057
4, 8, 16	–	0.455	0.069	96.399	1.019
MINLP	–	0.885	0.071	1.067	42.989
MILP	1	0.685	2.265	1.211	4.369
MILP	$2H$	0.704	0.081	1.117	0.592
MILP	100	0.112	0.000	17.148	0.386

- MIPs: used info. abt. type of gross error; EVM too expensive
- RE tailored to process $\Rightarrow$ Tune RE parameters to minimize $\mathcal{AIC}$

90

Conclusions – Data Reconciliation

- Redescending estimators superior for gross error detection and provide good parameter estimates
- Redescending estimators
  - More robust than Huber estimator: Fair function
  - Can be tuned
- Simple two-step tuning strategy utilized: considerably improves DRPE
- MINLP and MILP: good but computationally intensive and useful only for linearly constrained problems

91